ABSTRACT

A system and method for load balancing a plurality of servers is disclosed. In a preferred embodiment, a plurality of servers in a video-on-demand or other multi-server system are divided into one or more load-balancing groups. Each server preferably maintains state information concerning other servers in its load-balancing group including information concerning content maintained and served by each server in the group. Changes in a server's content status or other state information are preferably proactively delivered to other servers in the group. When a content request is received by any server in a load-balancing group, it evaluates the request in accordance with a specified algorithm to determine whether it should deliver the requested content itself or redirect the request to another server in its group. In a preferred embodiment, this determination is a function of information in the server's state table.